

Fast and memory-optimal dimension reduction using Kac’s walk

Vishesh Jain
Stanford University
vishesh.vj@gmail.com

Natesh S. Pillai
Harvard University
pillai@fas.harvard.edu

Ashwin Sah
Massachusetts Institute of Technology
asah@mit.edu

Mehtaab Sawhney
Massachusetts Institute of Technology
msawhney@mit.edu

Aaron Smith
University of Ottawa
asmi28@uottawa.ca

Abstract

In this work, we analyze dimension reduction algorithms based on the Kac walk and discrete variants.

- For n points in \mathbb{R}^d , we design an optimal Johnson-Lindenstrauss (JL) transform based on the Kac walk which can be applied to any vector in time $O(d \log d)$ for essentially the same restriction on n as in the best-known transforms due to Ailon and Liberty [SODA, 2008], and Bamberger and Kraemer [arXiv, 2017]. Our algorithm is memory-optimal, and outperforms existing algorithms in regimes when n is sufficiently large and the distortion parameter is sufficiently small. In particular, this confirms a conjecture of Ailon and Chazelle [STOC, 2006] in a stronger form.
- The same construction gives a simple transform with optimal Restricted Isometry Property (RIP) which can be applied in time $O(d \log d)$ for essentially the same range of sparsity as in the best-known such transform due to Ailon and Rauhut [Discrete Comput. Geom., 2014].
- We show that by fixing the angle in the Kac walk to be $\pi/4$ throughout, one obtains optimal JL and RIP transforms with almost the same running time, thereby confirming – up to a $\log \log d$ factor – a conjecture of Avron, Maymounkov, and Toledo [SIAM J. Sci. Comput., 2010]. Our moment-based analysis of this modification of the Kac walk may also be of independent interest.

1 Introduction

The aim of this paper is to design fast and simple dimensionality reduction algorithms with optimal embedding dimension – specifically, fast Johnson-Lindenstrauss (JL) transforms and fast Restricted Isometry Property (RIP) transforms – using the Kac walk and some of its discrete variants.

1.1 Fast Johnson-Lindenstrauss Transforms (FJLTs)

The classical lemma of Johnson and Lindenstrauss [20] asserts that for any collection of n points x_1, \dots, x_n in Euclidean space \mathbb{R}^d , and for any error parameter $\epsilon \in (0, 1)$, there exists a linear transformation $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}^k$, with $k = O(\epsilon^{-2} \log n)$, such that for all $i \in [n]$, $\|\Phi x_i\|_2 = (1 \pm \epsilon) \cdot \|x_i\|_2$. At least for $\epsilon > d^{-0.49}$, the bound on k is known to be optimal up to constants ([24, 5]).

The early examples of optimal JL embeddings (i.e. JL embeddings with asymptotically optimal embedding dimension) are (suitably rescaled) random Gaussian [15] and random Rademacher matrices [1]. While achieving the optimal embedding dimension for essentially all settings of the parameters d, n, ϵ , such embeddings are unfortunately too slow for many applications, since the time to compute the image Φv of a fixed vector $v \in \mathbb{R}^d$ is in general $O(dk) = O(\epsilon^{-2} d \log n)$.

To address this issue, optimal JL embeddings for which the image Φv of a fixed vector $v \in \mathbb{R}^d$ can be computed in time $O(d \log d)$ (we will often refer to this as the *running time*), under some restrictions on n and ϵ , have been proposed, starting with the seminal work of Ailon and Chazelle [2], who constructed a family of optimal JL embeddings with running time

$$O(d \log d + \min\{\epsilon^{-2} d \log n, \epsilon^{-2} \log^3 n\}). \quad (1.1)$$

In particular, for the $d \log d$ term to dominate the second term, we must have

$$n \leq \exp(\tilde{O}(\epsilon^{2/3} d^{1/3})), \quad (1.2)$$

where \tilde{O} hides possible logarithmic factors in ϵ and d . At least for fixed ϵ , this restriction was significantly relaxed by Ailon and Liberty [3], who provided a different family of optimal JL embeddings (for any $\gamma > 0$) with runtime

$$O(d \log(\epsilon^{-2} \log n)) \quad \text{for all } n \leq \exp(O_{\gamma, \epsilon}(d^{1/2-\gamma})). \quad (1.3)$$

In a recent work of Bamberger and Kraemer [8], an optimal JL embedding, which is simpler than the construction in [3], is provided with runtime

$$O(d \log(\epsilon^{-2} \log n)) \quad \text{for all } n \leq \exp(\tilde{O}(\epsilon^2 d^{1/2})). \quad (1.4)$$

Note that for the regime not covered by (1.2), the running time of $O(d \log(\epsilon^{-2} \log n))$ in (1.3) and (1.4) simplifies to $O(d \log d)$ as well.

Finally, we note that there is a separate line of work focused on designing optimal JL embeddings with even faster running times on sparse vectors; since this is not the focus of the present work, we omit further discussion, and refer the reader to [22], noting only that these sparse JL transforms may be used to improve the first term inside the min in (1.1) to $\epsilon^{-1} d \log n$.

1.2 The Restricted Isometry Property (RIP) and fast RIP transforms.

The design of JL embeddings with running time $O(d \log d)$ (albeit with suboptimal embedding dimension) for $n = \exp(\omega_\epsilon(\sqrt{d}))$ is based on the connection between JL transforms and transforms satisfying the Restricted Isometry Property (RIP). We recall this important notion, which was first isolated in the compressed sensing literature [11, 18].

Definition 1.1. For a matrix A , define

$$\delta_s(A) = \sup_{\substack{\|x\|_2=1 \\ x \text{ is } s\text{-sparse}}} |\|Ax\|_2^2 - 1|.$$

We say that A has the *Restricted Isometry Property (RIP) of order s and level δ* if $\delta_s(A) \leq \delta$.

Remark. It is known (see, e.g., [9]) that for any $k \times d$ matrix A , $\delta_s(A) \gtrsim \sqrt{(s \log(d/s))/k}$. Hence, we will informally say that a $k \times d$ matrix A is *RIP-optimal at s* if

$$\delta_s(A) \lesssim \sqrt{\frac{s \log(d/s)}{k}}. \quad (1.5)$$

As in the case of optimal JL transforms, the early constructions of optimal RIP transforms are based on random subgaussian matrices (see, e.g., [9]). Once again, these transforms have the drawback of not supporting fast matrix-vector multiplication, leading to the study of fast (nearly) optimal RIP transforms i.e. $k \times d$ matrices supporting matrix-vector multiplication in time $O(d \log d)$, and which satisfy the RIP property of order s and level δ with s, δ, k related (possibly up to polylogarithmic factors in d) as in (1.5).

Notably, improving on previous work of Candes and Tao [12], Rudelson and Vershynin [30] showed that a (suitably rescaled) random sample of $k = \Omega(\delta^{-2} s \log^4 d)$ rows of the Walsh-Hadamard matrix satisfies, with high probability, the RIP of order s and level δ . Since the Walsh-Hadamard matrix supports time $O(d \log d)$ matrix-vector multiplication via the Fast Walsh-Hadamard Transform, this gives a fast, nearly-optimal (in terms of embedding dimension) RIP transform. The result of Rudelson and Vershynin is optimal up to a factor of $\log^3 d$, which has since been improved (at least if one is willing to allow a slightly worse dependence on δ) – see [19] for an account of these developments.

There is a certain sense in which optimal RIP and optimal JL transforms are nearly equivalent. Indeed, an ϵ -net argument shows (see [9] for details) that optimal JL embeddings are also optimal RIP embeddings. In particular, this shows that the fast JL embedding in (1.3) gives a fast optimal RIP transform for

$$s \leq O_{\gamma, \delta}(d^{1/2-\gamma}).$$

In later work of Ailon and Rauhut [4], a simpler fast optimal RIP transform was obtained for

$$s \leq \tilde{O}(\delta d^{1/2}). \quad (1.6)$$

We also note that the optimal JL transform in (1.4) can be used to obtain an even simpler fast optimal RIP transform up to

$$s \leq \tilde{O}(\delta^2 d^{1/2}), \quad (1.7)$$

although this connection does not seem to have been observed in [8].

In the other direction, a remarkable result of Krahmer and Ward [23] (see also Theorem 3.5 below) shows that for any $k \times d$ matrix A with RIP of order s and level $\delta/4$, the random matrix AD , where D is a random diagonal Rademacher matrix, satisfies (with high probability) the JL property for a given collection of n points with error δ , provided that $n \leq 2^s$. This result will prove to be crucial for us.

1.3 The Kac walk and Orthogonal Repeated Averaging (ORA)

Introduced by Mark Kac [21] in 1956 as a toy model for a one-dimensional Boltzmann gas, the Kac walk is the following discrete time Markov chain $\{Q_t\}_{t \geq 0}$ on the special orthogonal group $\text{SO}(d)$.

Definition 1.2. Let $Q_0 = I_d \in \mathbb{R}^{d \times d}$. For all integers $t \geq 1$, sample two distinct uniform random coordinates $i_t, j_t \in [d]$ and a uniform random angle θ_t from $[0, 2\pi)$. Then, let $Q_t = R_{i_t, j_t, \theta_t} Q_{t-1}$, where $R_{i, j, \theta} \in \mathbb{R}^{d \times d}$ is the rotation in the (i, j) plane given by:

$$\begin{aligned} R_{i, j, \theta}(e_k) &= e_k \quad \text{for all } k \notin \{i, j\}; \\ R_{i, j, \theta}(x_i e_i + x_j e_j) &= (x_i \cos \theta - x_j \sin \theta)e_i + (x_i \sin \theta + x_j \cos \theta)e_j. \end{aligned}$$

By the *Kac walk of length T* we mean the random variable Q_T .

The Kac walk has a rich history in probability and mathematical physics (see, e.g., the references in [27, 28]). Its utility for dimensionality reduction was first suggested by Ailon and Chazelle [2], who also noted that the Kac walk has the attractive property that given the update sequence $\{R_{i_t, j_t, \theta_t}\}_{t \in [T]}$, the image $Q_T v$ of any vector $v \in \mathbb{R}^d$ can be computed with only a *constant* amount of memory overhead. Ailon and Chazelle conjectured that the Kac walk performs at least as well as fast JL transforms based on the fast Walsh-Hadamard transform. Specifically, they conjectured that for a given set of n points in \mathbb{R}^d and error parameter ϵ , projecting Q_T onto the first $O(\epsilon^{-2} \log n)$ coordinates gives a JL embedding of the point set with relative error ϵ , provided that

$$T = O(d \log d + \text{poly}(\log n, \epsilon^{-1})).$$

Recently, Choromanski, Rowland, Chen, and Weller [14] provided numerical support for this conjecture.

Despite this numerical evidence, the conjecture of Ailon and Chazelle may perhaps seem quite surprising from the point of view of mixing times of Markov chains. The initial proof of the JL lemma due to Johnson and Lindenstrauss [20] is based on taking the embedding matrix to be a uniformly random sample from the $O(\epsilon^{-2} \log n)$ -Stiefel manifold in d -dimensions (i.e. the uniform distribution over the set of $O(\epsilon^{-2} \log n)$ -orthonormal frames in \mathbb{R}^d). On the other hand, dimensional considerations show that the Kac walk *does not* mix on the k -Stiefel manifold in d -dimensions in $\Omega(kd)$ steps (see [26, Theorem 6] for a formal proof); in our setting, this would give a lower bound of $\Omega(\epsilon^{-2} d \log n)$, which asymptotically matches multiplication by a $O(\epsilon^{-2} \log n) \times d$ random Gaussian matrix. Indeed, Oliveira [26] conjectured that the Kac walk may be used to design JL transforms running in time $O(\epsilon^{-2} (\log n) d \log d)$ (this is slower than multiplication by a Gaussian matrix, but only requires a constant memory overhead) based on this connection with mixing on the $O(\epsilon^{-2} \log n)$ -Stiefel manifold in d -dimensions.

Nevertheless, as our first main result ([Theorem 1.4](#)), we confirm the conjecture of Ailon and Chazelle in a much stronger form by showing that the mixing of the Kac walk on the 1-Stiefel manifold in d -dimensions (i.e. the unit sphere) is already enough for the purpose of dimensionality reduction – while simple in hindsight, we believe that this provides a more intuitive and principled explanation for the existence of JL transforms running in time $O(d \log d)$ than is obtained from the analysis transforms based on Hadamard matrices. Specifically, we provide a fast and memory-optimal JL transform based on the Kac walk running in time

$$T = O(d \log d + \min\{d \log n, \epsilon^{-2} \log^2 n \log^2(\log n) \log^3 d\}).$$

In particular, the first term dominates the running time provided that

$$n \leq \exp(\tilde{O}(\epsilon d^{1/2})),$$

which matches the regime covered by (1.3) and (1.4) in terms of d (up to polylogarithmic factors), and has improved dependence on the error parameter ϵ (in particular, the first term inside the min is $d \log n$, which improves on the previously best known rate of $\epsilon^{-1} d \log n$ obtained by sparse JL transforms). As a corollary (Corollary 1.5), we also obtain a fast RIP transform, much simpler than in [4], up to

$$s \leq \tilde{O}(\delta d^{1/2}),$$

which matches the restriction in (1.6) up to polylogarithmic factors in d and is better in terms of δ dependence than (1.7).

For the purpose of computing matrix-vector products, an even faster and more elegant approach is to fix the angle in the Kac walk for the entire process to be $\theta = \pi/4$. This leads to the following discrete time Markov chain $\{Q_t\}_{t \geq 0}$ on $\text{SO}(d)$, which we call orthogonal repeated averaging (ORA) due to its apparent similarity to various iterated averaging processes in the probability literature (see, e.g., [13] and the references therein).

Definition 1.3. Let $Q_0 = I_d \in \mathbb{R}^{d \times d}$. For all integers $t \geq 1$, sample two distinct uniform random coordinates $i_t, j_t \in [d]$, and let $Q_t = R_{i_t, j_t} Q_{t-1}$, where $R_{i, j} \in \mathbb{R}^{d \times d}$ is the rotation in the (i, j) plane given by:

$$\begin{aligned} R_{i, j, \theta}(e_k) &= e_k \quad \text{for all } k \notin \{i, j\}; \\ R_{i, j, \theta}(x_i e_i + x_j e_j) &= \left(\frac{x_i + x_j}{\sqrt{2}} \right) e_i + \left(\frac{x_i - x_j}{\sqrt{2}} \right) e_j. \end{aligned}$$

By the *ORA of length T* we mean the random variable Q_T .

The application of ORA to dimensionality reduction was suggested by Avron, Maymounkov, and Toledo [6], who conjectured (based on experimental evidence) that the ORA performs as well as the Kac walk for dimensionality reduction. From the point of view of mixing times, this is even more delicate since, for instance, the total variation distance between the uniform distribution on the sphere and the ORA distribution for *any* finite number of steps is always 1. In our final main result (Theorem 1.6), we almost confirm this conjecture by designing an optimal JL transform based on ORA running in time

$$T = O(d \log d \log \log d + \log d \min\{d \log n, \epsilon^{-2} \log^2 n \log^2(\log n) \log^3 d\}).$$

As in Corollary 1.5, this also gives an RIP-optimal transform.

Remark. We conjecture that the additional $\log \log d$ factor (which is anyway essentially constant for practical purposes) can be removed, and expect the ORA based transform to be more efficient than the Kac walk based transformed in practice.

We now proceed to a formal statement of our main results.

1.4 Main Results

As mentioned above, our first result is a fast JL transform based on the Kac walk which essentially matches the fastest known JL transforms based on subsampled Hadamard matrices in all regimes, and improves all known transforms in some regimes.

Theorem 1.4. *There is an absolute constant $C_{1.4} > 0$ for which the following holds. Let $d, n, T, \epsilon > 0$ satisfy $n \geq d \geq C_{1.4}$, $\epsilon \in (0, C_{1.4}^{-1})$, and $\epsilon^{-2} \log n \leq d$. Then, [Algorithm 1](#) runs in time*

$$T \leq C_{1.4} \left(d \log d + \min \{ d \log n, \epsilon^{-2} (\log n)^2 (\log \log n)^2 (\log d)^3 \} \right),$$

and outputs a linear map $\Psi : \mathbb{R}^d \rightarrow \mathbb{R}^k$, where $k = k(n, \epsilon) \leq C_{1.4} \epsilon^{-2} \log n$, such that for any fixed set $X \subseteq \mathbb{R}^d$ of size $|X| = n$, the inequalities

$$(1 - \epsilon) \|x\|_2 \leq \|\Psi x\|_2 \leq (1 + \epsilon) \|x\|_2$$

hold simultaneously for all $x \in X$ with probability at least $2/3$.

Furthermore, for any $x \in \mathbb{R}^d$, Ψx can be computed in time $O(T)$ with only $O(1)$ additional memory.

Remark. The probability of failure can be improved to $1 - \eta$ with easy and standard modifications of the proof. For the sake of simplicity, we do not keep track of the dependence on η . Also, the restriction $\epsilon^{-2} \log n \leq d$ is of no significance since when $d \leq \epsilon^{-2} \log n$, one may simply take ψ to be the identity map.

As a direct corollary, we obtain a much simpler construction than in [4] of a fast RIP-optimal transformation for s matching the restriction in (1.6).

Corollary 1.5. *Fix $\eta > 0$. Given $\delta \in (0, C_{1.5}^{-1})$, d , and $s \leq \tilde{O}(\delta d^{1/2})$, let $k = C_{1.5} \epsilon^{-2} s \log d$. Define Ψ as in [Algorithm 1](#), where $n = d^s (1 + 2/\delta)^s / s!$. Then with probability at least $2/3$, Ψ is RIP of order s and level $C_{1.5} \delta$ and has image dimension at most k . Furthermore, application of Ψ on a given point takes $O(d \log d)$ time.*

Remark. Unlike approaches based on the Fast Walsh-Hadamard Transform (e.g. [2, 4, 8]), [Algorithm 1](#) requires neither any preconditioning by random signed diagonal matrices, nor any postconditioning by random permutation matrices.

Finally, we obtain similar results, even after replacing the Kac walk by the simpler ORA.

Theorem 1.6. *There is an absolute constant $C_{1.6} > 0$ for which the following holds. Let $d, n, T, \epsilon > 0$ satisfy $n \geq d \geq C_{1.6}$, $\epsilon \in (0, C_{1.6}^{-1})$, and $\epsilon^{-2} \log n \leq d$. Then, [Algorithm 2](#) runs in time*

$$T \leq C_{1.6} \left(d \log d \log \log d + \min \{ d \log d \log n, \epsilon^{-2} (\log n)^2 (\log \log n)^2 (\log d)^4 \} \right),$$

and outputs a linear map $\Psi : \mathbb{R}^d \rightarrow \mathbb{R}^k$, where $k = k(n, \epsilon) \leq C_{1.6} \epsilon^{-2} \log n$, such that for any fixed set $X \subseteq \mathbb{R}^d$ of size $|X| = n$, the inequalities

$$(1 - \epsilon) \|x\|_2 \leq \|\Psi x\|_2 \leq (1 + \epsilon) \|x\|_2$$

hold simultaneously for all $x \in X$ with probability at least $2/3$.

Furthermore, for any $x \in \mathbb{R}^d$, Ψx can be computed in time $O(T)$ with only $O(1)$ additional memory.

1.5 Techniques

Our algorithms are very simple, and are best viewed as running in two phases: in the first phase, we achieve an embedding into a nearly optimal dimension, and in the second phase, we correct this nearly optimal dimension to the optimal dimension. Despite the simplicity of the algorithms, the analysis is involved, and makes use of a multitude of techniques from probability and high-dimensional geometry.

The analysis of the first phase – which is simply either the Kac walk on \mathbb{R}^d (with no preconditioning), or the ORA on \mathbb{R}^d preceded by a single preconditioning step – boils down to two things. First, we need the near-optimal JL property of randomly subsampled rows of ‘bounded’ orthogonal matrices, which is proved via chaining methods (see [Theorem 3.4](#) and [Theorem 3.5](#)). Second, we need to show that $O(d \log d)$ steps of the Kac walk or ORA lead to sufficiently bounded orthogonal systems. For the Kac walk, this follows by making use of a contractive coupling, introduced in work of Pillai and Smith [\[27\]](#) studying the total variation mixing time of the Kac walk on the sphere. Unfortunately, for the ORA, this coupling breaks down (as noted earlier, after any finite number of steps, the ORA distribution has total variation distance 1 from the uniform distribution on the sphere, which explains some of the difficulty in devising coupling-based arguments); we get around this by a completely different argument based on combining the FKG correlation inequality with a delicate recursive computation of various (weighted) moments of the ORA distribution. Furthermore, we demonstrate various symmetry properties of the Kac walk, which enable us to forego the preconditioning by random signed diagonal matrices and/or postconditioning by random permutation matrices present in previous works in this area [\[2, 3, 8, 4\]](#) – this involves bringing in tools from nonabelian Fourier analysis, in particular adapting work of Diaconis and Shahshahani [\[16\]](#) on the transposition walk on the symmetric group \mathfrak{S}_d .

The second phase of our algorithm is either a Kac walk on $\mathbb{R}^{d'}$ or ORA on $\mathbb{R}^{d'}$, run for $O(d' \log n)$ steps, where d' is the intermediate nearly-optimal dimension from the first step. For the Kac walk, the contractive coupling suffices for the analysis, whereas for the ORA, we need an analysis based on combining our moment computations with a theorem of Latała on the moments of sums of independent random variables.

We note that a two stage algorithm achieving similar objectives also appears in a recent work of Bamberger and Kraemer [\[8\]](#), although the two stages used in their work are very different from each other; they use randomly subsampled Hadamard matrices in the first stage, and random Gaussian matrices in the second stage, which leads to an error term of $\epsilon^{-2} d \log n$. Apart from achieving a better error term of $d \log n$, and being much more memory efficient, we additionally show how both of these two seemingly disparate stages can be accomplished by the same process (either Kac walk or ORA).

Finally, we remark that our analysis of ORA may be of independent interest (for instance, in probability and quantum computing, see [\[13\]](#)).

1.6 Organization

The rest of this paper is organized as follows. In [Section 3](#), we present and analyse our Kac-walk based algorithm (one of the proofs, present in [\[27\]](#), is included in [Appendix A](#) for completeness) and in [Section 4](#), we present and analyse our ORA-based algorithm. [Section 2](#) contains some auxiliary results related to removing various preconditioning and postconditioning operations; the proof of one of these results is contained in [Appendix B](#). Finally, [Section 5](#) contains some open problems and directions for future research.

2 Preliminaries

2.1 Projection and sampling operators

Let Proj_X be the projection operator from \mathbb{R}^m onto the first X basis vectors; sometimes, X will be a random variable, in which case this is to be understood as first generating X , and then projecting onto the first X basis vectors. Let $\text{Proj}_{m,q}$ be a (random) projection from \mathbb{R}^m to a random subset of basis vectors, where each is kept with probability q . Finally, let Sample_K be a (random) projection to a uniformly random subset of size K of the standard basis vectors. Given a random vector $\xi \in \{\pm 1\}^m$, let D_ξ denote the corresponding diagonal matrix. Note that all of these projections can trivially be computed in $O(m)$ time with $O(1)$ additional space.

2.2 Symmetric Kac walks

We now isolate the notion of symmetric random walks, which do not require any preconditioning by random diagonal Rademacher matrices or postconditioning by random permutation matrices. We begin by noting that both the standard Kac walk as well as ORA are instances of the following more general process.

Definition 2.1. For a distribution q on angles $[0, 2\pi)$, we define the q -Kac walk on $\text{SO}(d)$ as follows. Let $Q_0 = I_d \in \mathbb{R}^{d \times d}$. For all integers $t \geq 1$ sample two distinct uniform random coordinates $i_t, j_t \in [d]$ and a random angle θ_t from q , and let $Q_t = R_{i_t, j_t, \theta_t} Q_{t-1}$, where $R_{i, j, \theta} \in \mathbb{R}^{d \times d}$ is the rotation in the (i, j) plane given by:

$$\begin{aligned} R_{i, j, \theta}(e_k) &= e_k \text{ for } k \notin \{i, j\} \\ R_{i, j, \theta}(x_i e_i + x_j e_j) &= (x_i \cos \theta - x_j \sin \theta)e_i + (x_i \sin \theta + x_j \cos \theta)e_j. \end{aligned}$$

By the q -Kac walk of length T we mean the random variable Q_T .

Note that the standard Kac walk corresponds to q being the uniform distribution on $[0, 2\pi)$, and ORA corresponds to q being the delta distribution concentrated at $\pi/4$. We will refer to the standard Kac walk as simply the Kac walk.

Definition 2.2. A q -Kac walk is said to be symmetric if the distribution q is invariant under the maps $\theta \mapsto -\theta$ and $\theta \mapsto \theta + \pi/2$.

Clearly, the Kac walk is symmetric. ORA is not symmetric; however, taking q to be the uniform measure on $\{\pi/4, 3\pi/4, 5\pi/4, 7\pi/4\}$ leads to a symmetric walk, which we call symmetric ORA (S-ORA for short).

The following two lemmas about symmetric q -Kac walks enable us to dispense with various preconditioning/postconditioning operations appearing in the literature. (e.g., in [2, 4, 8, 3]).

Lemma 2.3. Consider a uniform vector $\xi \in \{\pm 1\}^d$, conditioned on having product 1. Then, for any symmetric q -Kac walk,

$$\text{TV}(Q_T, Q_T D_\xi) \leq \frac{d \exp\left(-\frac{T}{d-1}\right)}{1 - d \exp\left(-\frac{T}{d-1}\right)},$$

where TV denotes the total variation distance.

Proof. For every pair of distinct indices $i, j \in [d]$, let $D_{i,j}$ be the random diagonal matrix with all 1s, except in the positions i, j , where the entries are either both 1 or both -1 with equal probability. For every time $t \in [T]$, let D_t be a random diagonal matrix distributed as D_{i_t, j_t} , all sampled independently from everything except (i_t, j_t) .

First, note that R_{i_t, j_t, θ_t} and $R_{i_t, j_t, \theta_t} D_t$ have the same distribution since our distribution q on angles is invariant under $\theta \leftrightarrow \theta + \pi$. Second, note that the distributions of

$$D_{i', j'} R_{i, j, \theta} \text{ and } R_{i, j, \theta} D_{i', j'}$$

are the same (these being independent random matrices), using the symmetry $\theta \leftrightarrow -\theta$ and

$$D_{i', j'} R_{i, j, \theta} D_{i', j'}^{-1} = R_{i, j, -\theta}$$

in the case when $|\{i, j\} \cap \{i', j'\}| = 1$.

By applying the first operation to R_{i_T, j_T, θ_T} , and then applying the second operation repeatedly to switch the diagonal matrix D_T to the end, we see that Q_T has the same distribution as $Q_T D_T$. Then we do the same with $R_{i_{T-1}, j_{T-1}, \theta_{T-1}}$, and so on, and thus we have the same distribution as $Q_T D_{T-1} D_T$, and so on, until

$$Q_T D_1 \cdots D_T.$$

Now, let \mathcal{E} be the event that the graph on vertex set $[d]$ spanned by the edges (i_t, j_t) for $t \in [T]$ is connected. Condition on any instantiation of all the pairs (i_t, j_t) such that \mathcal{E} holds. We easily see that $D_1 \cdots D_T$ and D_ξ have the same distribution in this case, and furthermore that $Q_T D_1 \cdots D_T$ and $Q_T D_\xi$ also have the same distribution in this case (since after conditioning on our instantiation, Q_T is independent from D_t for $t \in [T]$ as well as D_ξ).

Therefore, conditional on \mathcal{E} , we have that $Q_T D_1 \cdots D_T$ and $Q_T D_\xi$ have the same distribution, so that

$$\text{TV}(Q_T, Q_T D_\xi) = \text{TV}(Q_T D_1 \cdots D_T, Q_T D_\xi) \leq \mathbb{P}[\mathcal{E}^c].$$

Finally, we have

$$\mathbb{P}[\mathcal{E}^c] \leq \frac{d \exp\left(-\frac{T}{d-1}\right)}{1 - d \exp\left(-\frac{T}{d-1}\right)};$$

this follows from well known results about the $O((\log d)/d)$ threshold for random graphs to be connected [10, Chapter 7]. \square

Remark. The true cutoff for connectedness occurs at $T = d \log d/2$ and not $T = d \log d$. However, deriving an exact expression suitable for non-asymptotic analysis is nontrivial, and is anyway not a crucial point in our final analysis.

In fact, as the next lemma shows, symmetric q -Kac walks enjoy a more non-trivial invariance property. Namely, after $O(d \log d)$ steps, the distribution is essentially invariant under left-multiplication by signed permutation matrices in $\text{SO}(d)$. This allows us to simplify our transforms further by simply projecting onto an initial segment of coordinates, thus enabling a more straightforward memory-optimal, in-place implementation.

Lemma 2.4. *Fix $d \geq 10$. Let Σ be a uniformly chosen signed permutation matrix in $\text{SO}(d)$ and D_ξ be as in Lemma 2.3. Then, for any symmetric q -Kac walk,*

$$\text{TV}(Q_T, \Sigma Q_T D_\xi) \leq \frac{2d \exp\left(-\frac{T}{d-1}\right)}{1 - d \exp\left(-\frac{T}{d-1}\right)} + C_{2.4} \left(d^{1/2} e^{-T/(6d)} + (d!)^{1/2} \left(\frac{\sqrt{5}-1}{2} \right)^{T/2} \right),$$

where $C_{2.4} > 0$ is an absolute constant.

The proof of this result is presented in Appendix B, and relies on character estimates of Diaconis and Shahshahani [16] used to prove a sharp cutoff for the transposition walk on \mathfrak{S}_d .

3 Fast JL-Optimal and RIP-Optimal Transforms Using the Kac Walk: Proof of [Theorem 1.4](#) and [Corollary 1.5](#)

The proof of [Theorem 1.4](#) and [Corollary 1.5](#) uses [Algorithm 1](#).

Algorithm 1: Fast JL via the Uniform Kac walk

#Run the uniform Kac walk for $O(d \log d)$ steps

Take $T_1 = 12d \log d$ and $K_1 = \min(d, C_1 \epsilon^{-2} \log n (\log \log n)^2 (\log d)^3)$. Sample Q_{T_1} from the uniform Kac walk and let

$$\Psi_1 := \sqrt{\frac{d}{K_1}} \cdot \text{Proj}_{\text{Binom}(d, K_1/d)} \circ Q_{T_1}.$$

Take $T_2 \geq 12K_1 \log n$ and $K_2 \geq C_1 \epsilon^{-2} \log n$. Sample Q'_{T_2} from the uniform Kac walk and let

$$\Psi_2 := \sqrt{\frac{K_1}{K_2}} \cdot \text{Proj}_{\text{Binom}(K_1, K_2/K_1)} \circ Q'_{T_2}.$$

Return

$$\Psi = \Psi_2 \circ \Psi_1.$$

3.1 Coupling and contraction estimates for the Kac walk on \mathbb{S}^{d-1}

In this subsection, we describe a coupling of two copies of the Kac walk X_t, Y_t so that the distance between them goes to zero exponentially quickly – this is one of the two key steps in our analysis of [Algorithm 1](#). To begin, note that the Kac walk may be viewed as a discrete-time Markov chain $\{X_t\}_{t \geq 0}$ on \mathbb{S}^{d-1} defined as follows: at every step t , choose two coordinates $1 \leq i_t < j_t \leq d$ and an angle $\theta_t \in [0, 2\pi)$ uniformly at random, and set

$$\begin{aligned} X_{t+1}[i_t] &= \cos(\theta_t) X_t[i_t] - \sin(\theta_t) X_t[j_t] \\ X_{t+1}[j_t] &= \sin(\theta_t) X_t[i_t] + \cos(\theta_t) X_t[j_t] \\ X_{t+1}[k] &= X_t[k] \quad k \notin \{i_t, j_t\}. \end{aligned} \tag{3.1}$$

Let $F : [d] \times [d] \times [0, 2\pi) \times \mathbb{S}^{d-1} \rightarrow \mathbb{S}^{d-1}$ be the map associated with this representation, so that $X_{t+1} = F(i_t, j_t, \theta_t, X_t)$.

Definition 3.1 (Proportional coupling, see Definition 3.1 in [\[27\]](#)). Define a coupling of two copies $\{X_t\}_{t \geq 0}, \{Y_t\}_{t \geq 0}$ of Kac's walk as follows. Fix $X_0, Y_0 \in \mathbb{S}^{d-1}$. Let (i_0, j_0, θ_0) be the update variables used by X_1 in [\(3.1\)](#). Choose $\varphi \in [0, 2\pi)$ uniformly at random among all angles that satisfy

$$\begin{aligned} X_1[i_0] &= \sqrt{X_0[i_0]^2 + X_0[j_0]^2} \cos \varphi, \\ X_1[j_0] &= \sqrt{X_0[i_0]^2 + X_0[j_0]^2} \sin \varphi. \end{aligned}$$

As noted in [\[27\]](#), if $X_0[i_0] = X_0[j_0] = 0$, then all angles φ satisfy this equation; otherwise, there is a unique such φ , and the value of $\varphi - \theta_0 \pmod{2\pi}$ does not depend on θ_0 .

Then, choose $\theta'_0 \in [0, 2\pi)$ uniformly among the angles that satisfy

$$F(i_0, j_0, \theta'_0, Y_0)[i_0] = \sqrt{Y_0[i_0]^2 + Y_0[j_0]^2} \cos \varphi,$$

$$F(i_0, j_0, \theta'_0, Y_0)[j_0] = \sqrt{Y_0[i_0]^2 + Y_0[j_0]^2} \sin \varphi,$$

and set $Y_1 = F(i_0, j_0, \theta'_0, Y_0)$. Note that this coupling forces Y_1 to be as close as possible to X_1 in the Euclidean distance (for instance, in two dimensions, we always have $X_1 = Y_1$ under this coupling, and in more than two dimensions, it still forces the points $(0, 0), (X_1[i_0], X_1[j_0]), (Y_1[i_0], Y_1[j_0])$ to be collinear).

Now, continue this process starting from (X_1, Y_1) instead of (X_0, Y_0) .

The following key lemma shows that, under the coupling described above, the distance (interpreted suitably) between two copies of Kac's walk decreases exponentially fast.

Lemma 3.2 (See Lemma 3.3 in [27]). *Fix $X_0, Y_0 \in \mathbb{S}^{d-1}$. For $t \geq 0$, couple (X_{t+1}, Y_{t+1}) conditional on (X_t, Y_t) according to the coupling in Definition 3.1. Then, for any $t \geq 0$, Kac's walk on \mathbb{S}^{d-1} satisfies*

$$\mathbb{E} \left[\sum_{i=1}^d (X_t[i]^2 - Y_t[i]^2)^2 \right] \leq 2 \left(1 - \frac{1}{2d} \right)^t \leq 2e^{-t/(2d)}.$$

For the reader's convenience, we include the complete (short) proof of this lemma in Appendix A. Given this contractive coupling, we now derive estimates regarding the boundedness of the coordinates of the Kac walk.

Lemma 3.3. *Fix $X_0 \in \mathbb{S}^{d-1}$. Then, for any $\epsilon \in (1/d, 1/2)$, any $K \geq 2$, any $k \in [d]$, and any $t \geq 0$, the (uniform) Kac walk satisfies the following, denoting $X_t = Q_t X_0$.*

1. $\mathbb{P} \left[\sum_{i=1}^k X_t[i]^2 \notin \frac{k}{d} [1 - \epsilon, 1 + \epsilon] \right] \leq 8d^4 \exp(-t/(2d)) + 2 \exp(-\epsilon^2 k/64);$
2. $\mathbb{P} \left[\max_{i,j \in [d]} |Q_t[i, j]| \geq K \sqrt{\frac{\log d}{d}} \right] \leq 2d^3 \exp(-t/(2d)) + 2d^{5/2} \exp(-K^2(\log d)/2).$

Proof. Let Y_0 be a uniformly sampled from the sphere \mathbb{S}^{d-1} and couple our Kac walk X_t (via the proportional coupling Definition 3.1) to a Kac walk Y_t starting from Y_0 . Then, we have

$$\mathbb{E} \left[\left| \sum_{i=1}^k (X_t[i]^2 - Y_t[i]^2) \right|^2 \right] \leq k \mathbb{E} \left[\sum_{i=1}^d (X_t[i]^2 - Y_t[i]^2)^2 \right] \leq 2k \left(1 - \frac{1}{2d} \right)^t \leq 2ke^{-t/(2d)}$$

by Lemma 3.2 and Cauchy–Schwarz. Therefore, Markov's inequality implies that

$$\mathbb{P} \left[\left| \sum_{i=1}^k (X_t[i]^2 - Y_t[i]^2) \right| \geq \epsilon k/(2d) \right] \leq 8d^2 k^{-1} \epsilon^{-2} e^{-t/(2d)} \leq 8d^4 e^{-t/(2d)}. \quad (3.2)$$

Given this, it suffices to show that $\sum_{i=1}^k Y_t[i]^2$ is well-concentrated, which follows since Y_t is uniformly distributed on \mathbb{S}^{d-1} ; we include a short computation demonstrating this well-known fact for completeness.

Let c be a constant to be specified later. Let Z and Z' be independent uniform random vectors on \mathbb{S}^{d-1} . Then,

$$\begin{aligned} \mathbb{P} \left[\left| \sum_{i=1}^k Z[i]^2 - \frac{k}{d} \right| \geq \frac{\epsilon k}{2d} \right] &\leq e^{-\frac{c\epsilon k}{2d}} \mathbb{E} \left[e^{c(\sum_{i=1}^k Z[i]^2 - \frac{k}{d})} + e^{-c(\sum_{i=1}^k Z[i]^2 - \frac{k}{d})} \right] \\ &\leq e^{-\frac{c\epsilon k}{2d}} \mathbb{E} \left[e^{c(\sum_{i=1}^k Z[i]^2 - Z'[i]^2)} + e^{-c(\sum_{i=1}^k Z[i]^2 - Z'[i]^2)} \right] \end{aligned}$$

$$= 2e^{-\frac{c\epsilon k}{2d}} \mathbb{E} \left[e^{c(\sum_{i=1}^k Z[i]^2 - Z'[i]^2)} \right],$$

where in the second line we have used Jensen's inequality and in the third line we have used symmetry. Let $r = \sum_{i=1}^d G[i]^2$ where $G[i] \sim \mathcal{N}(0, 1/d)$ and let r' be an independent copy of r . Using orthogonal invariance of the Gaussian we have that $r \cdot Z$ (i.e. we pointwise multiply each coordinate of Z by r) is distributed as a Gaussian vector $(G[1], \dots, G[d])$ with each coordinate distributed as $\mathcal{N}(0, 1/d)$. Using these properties along with $\mathbb{E}[r] = \mathbb{E}[r'] = 1$, we see that

$$\begin{aligned} 2e^{-\frac{c\epsilon k}{2d}} \mathbb{E} \left[e^{c(\sum_{i=1}^k Z[i]^2 - Z'[i]^2)} \right] &\leq 2e^{-\frac{c\epsilon k}{2d}} \mathbb{E} \left[e^{c(\sum_{i=1}^k G[i]^2 - G'[i]^2)} \right] \\ &= 2e^{-\frac{c\epsilon k}{2d}} \mathbb{E} \left[e^{cG^2} \right]^k \mathbb{E} \left[e^{-cG^2} \right]^k \\ &= 2e^{-\frac{c\epsilon k}{2d}} (1 - 4c^2/d^2)^{-k/2}, \end{aligned}$$

where we have use Jensen's inequality to replace $Z[i]^2$ by $(r \cdot Z)[i]^2$ and $Z'[i]^2$ by $(r' \cdot Z')[i]^2$, then independence between coordinates, and then explicit computation (assuming $c < d/2$). Now let $c = A\epsilon d$, so that ultimately

$$\mathbb{P} \left[\left| \sum_{i=1}^k Z[i]^2 - \frac{k}{d} \right| \geq \frac{\epsilon k}{2d} \right] \leq 2e^{-A\epsilon^2 k/2} (1 - 4A^2 \epsilon^2)^{-k/2} \leq 2e^{-A\epsilon^2 k/2 + 4A^2 \epsilon^2 k}.$$

Finally, letting $A = 1/16$ and union-bounding with (3.2) proves conclusion 1. of the lemma.

For the second conclusion, it suffices to prove that

$$\mathbb{P} \left[\max_i |X_t[i]| \geq K \sqrt{\frac{\log d}{d}} \right] \leq 2d^2 \exp(-t/(2d)) + 2d^{3/2} \exp(-K^2(\log d)/2),$$

since then, union bounding over $X_0 = e_1, \dots, e_d$ immediately gives the desired result.

For this, we that Markov's inequality combined with Lemma 3.2 gives

$$\mathbb{P} \left[\max_i |X_t[i]^2 - Y_t[i]^2| \geq \frac{\log d}{d} \right] \leq 2d^2 e^{-t/(2d)}. \quad (3.3)$$

Since Y_t is uniformly distributed on the sphere, we have good control over $\max_i |Y_t[i]|^2$. In particular, recall a standard bound on the volume of spherical caps (see e.g., [7, Lemma 2.2]): for a uniformly random unit vector $Y_t \in \mathbb{R}^d$ and a basis vector $e_i \in \mathbb{R}^d$, we have

$$\mathbb{P}[Y_t[i] \geq \epsilon] = \mathbb{P}[|Y_t - e_i|^2 \leq 2 - 2\epsilon] \leq e^{-d\epsilon^2/2}. \quad (3.4)$$

Similarly, one obtains the same bound for $\mathbb{P}[Y_t[i] \leq -\epsilon]$. Using these two bounds with $\epsilon = \sqrt{(K^2 - 1)(\log d)/d}$ and taking the union bound over $1 \leq i \leq d$, we see that

$$\mathbb{P} \left[\max_i Y_t[i]^2 \geq (K^2 - 1) \frac{\log d}{d} \right] \leq 2de^{-(K^2 - 1)(\log d)/2} = 2d^{3/2} e^{-K^2(\log d)/2},$$

which combined with (3.3) gives the desired result. \square

3.2 JL-optimality

The other key tool in proving [Theorem 1.4](#) is a (by now) classic result [[12](#), [30](#), [29](#)] that demonstrates the restricted isometry property of orthogonal matrices with ‘bounded’ coordinates. We cite the version due to Dirksen [[17](#)], which provides the best known bounds if one requires the dependence on δ to be optimal i.e. δ^{-2} .

Theorem 3.4 ([[17](#), Theorem 4.1]). *Let U be an $N \times N$ orthogonal matrix with $\sup_{i,j \in [N]} \sqrt{N} |U_{i,j}| \leq K$. Recall δ_s is defined as*

$$\delta_s(A) = \sup_{\substack{\|x\|=1 \\ s\text{-sparse}}} \left| \|Ax\|_2^2 - 1 \right|.$$

Then, $\mathbb{P}[\delta_s(U_I) \geq \delta] \leq \eta$, where $U_I = \sqrt{N/m} \cdot \text{Proj}_{N,q} \circ U$ and $q = m/N$, as long as

$$m \geq C_{3.4} s K^2 \delta^{-2} \max((\log s)^2 (\log m) (\log N), \log(\eta^{-1})).$$

Finally, we need the following slight modification of the previously mentioned result of Krahmer and Ward [[23](#)] which, along with [Lemma 2.3](#), will allow us to deduce a Johnson-Lindenstrauss property based on the restricted isometry property of the uniform Kac walk.

Theorem 3.5 (Modified [[23](#), Theorem 3.1]). *Fix $\eta > 0$ and $\epsilon \in (0, 1)$, and consider a finite set $E \subseteq \mathbb{R}^d$ of cardinality $|E| = n$. Set $k \geq C_{3.5} \log(4n/\eta)$, and suppose that $\Phi \in \mathbb{R}^{m \times d}$ satisfies the Restricted Isometry Property of order k and level $\delta \leq \epsilon/4$. Let $\xi \in \mathbb{R}^d$ be a uniform vector in $\{\pm 1\}^d$, conditioned on having product 1, and let D_ξ denote the $d \times d$ diagonal matrix whose diagonal entries are given by ξ . Then, with probability at least $1 - \eta$,*

$$(1 - \epsilon) \|x\|_2^2 \leq \|\Phi D_\xi x\|_2^2 \leq (1 + \epsilon) \|x\|_2^2$$

uniformly for all $x \in E$.

Proof sketch. The proof is identical to the one given in [[23](#)] once we note that the proof in [[23](#)] only requires that the vector ξ is distributed as an independent Rademacher vector when restricted to certain proper subsets of $[d]$, which this altered random variable clearly satisfies. \square

We now have all the tools needed to prove [Theorem 1.4](#).

Proof of [Theorem 1.4](#). Let $q = K_1/d$. Applying [Theorem 3.4](#) and using the second part of [Lemma 3.3](#) at time $t = T_1 = 12d \log d$, we see that

$$\Psi'_1 = \frac{1}{\sqrt{q}} \text{Proj}_{d,q} \circ Q_{T_1},$$

with probability $1 - O(1/d)$, satisfies $\mathbb{P}[\delta_s(\Psi'_1) \geq \epsilon/4] \leq 1/d$ as long as

$$K_1 \gtrsim s (\log d) \epsilon^{-2} (\log s)^2 (\log K_1) (\log d).$$

Note in the case $K_1 = d$, the operator Ψ'_1 is actually orthogonal.

Now by [Theorem 3.5](#), we have that if $\delta_s(\Psi'_1) \leq \epsilon/4$ and $s \geq 40 \log(4n/\eta)$, then $\Psi'_1 \circ D_\xi$ acts as a $(1 \pm \epsilon)$ -isometry on our set of points X with probability at least $1 - \eta$. Choosing $\eta = 1/4$ and $s = 40 \log(16n)$, we see that this property holds with probability at least $3/4 - O(1/d)$ as long as

$$K_1 \gtrsim \epsilon^{-2} (\log n) (\log d)^2 (\log \log n)^2 (\log K_1).$$

Since $K_1 \leq d$, certainly

$$K_1 \gtrsim \epsilon^{-2}(\log n)(\log d)^3(\log \log n)^2$$

suffices. It is easily seen that as long as C_1 is chosen large enough, this bound indeed holds for the choice of K_1 in [Algorithm 1](#).

By Markov's inequality, the probability that the actual number of dimensions in the image of $\text{Proj}_{d,q}$ is more than a constant times K_1 is sufficiently small, so with probability at least $5/7$, we have that $\Psi'_1 \circ D_\xi$ is a $(1 \pm \epsilon)$ -isometry on our points and projects down to at most $O(\epsilon^{-2}(\log n)(\log d)^3(\log \log n)^2)$ dimensions.

Next, we show that Ψ_2 is a $(1 \pm \epsilon)$ -isometry on the image of our point set, $(\Psi'_1 \circ D_\xi)X$. In particular, applying the union bound using the first part of [Lemma 3.3](#) over all n vectors in the image immediately gives the desired result as long as C_1 is large enough. Since $\Psi'_1 \circ D_\xi, \Psi_2$ are both $(1 \pm \epsilon)$ -isometries on the relevant sets of points, it follows (after rescaling ϵ) that the composition satisfies the desired isometry property with probability at least, say, $7/10$.

Finally, note that in [Algorithm 1](#), we use Ψ_1 instead of the more complicated $\Psi'_1 \circ D_\xi$ – that this can be done follows easily from [Lemma 2.4](#) (and after decreasing the probability of success slightly from $7/10$ to say, $2/3$).

We now quickly compute the runtime and memory of [Algorithm 1](#). In order to compute $\Psi_i x$ (for $i = 1, 2$), we apply rotations R_{i_t, j_t, θ_t} in sequence, and then sparsify. This clearly requires constant memory as computations can be done in place, and since each R_{i_t, j_t, θ_t} affects at most 2 coordinates at once, the runtime is $O(d \log d + K_1 \log n)$. \square

3.3 RIP-optimality

The proof of [Corollary 1.5](#) follows exactly as in [9, Lemma 5.1].

Proof sketch of [Corollary 1.5](#). This is an application of [Theorem 1.4](#), noting that the size of a δ -net of all s -sparse unit vectors in \mathbb{R}^d is at most $\binom{d}{s}(1 + 2/\delta)^s$. \square

4 Fast JL-Optimal and RIP-Optimal Transforms Using ORA: Proof of [Theorem 1.6](#)

The proof of [Theorem 1.6](#) uses [Algorithm 2](#).

The analysis of [Algorithm 2](#) follows the same high level outline as the analysis of [Algorithm 1](#). However, due to the unavailability of a tractable contractive coupling between the ORA and uniform distribution on the sphere, the proof of the analogues of [Lemma 3.3](#) is more intricate. We now proceed to the details.

Definition 4.1. For a vector $\mathbf{x} \in \mathbb{R}^d$ and for $k \in \mathbb{N} \cup \{0\}$, define

$$S_k(\mathbf{x}) := \frac{1}{(2k)!} \sum_{i=1}^d x_i^{2k}.$$

In particular, $S_0(\mathbf{x}) = d$.

The next simple but crucial lemma studies the evolution of $S_k(\mathbf{x})$ under one step of ORA.

Algorithm 2: Fast JL via ORA

#Run orthogonal repeated averaging for $O(d \log d)$ steps

Take $T_1 = C_{4.4} d \log d \log \log d$ and $K_1 = \min(d, C_2 \epsilon^{-2} \log n (\log \log n)^2 (\log d)^3)$. Sample Q_{T_1} from ORA and D an independent diagonal random Rademacher matrix, and let

$$\Psi_1 := \sqrt{\frac{d}{K_1}} \cdot \text{Proj}_{d, K_1/d} \circ Q_{T_1} \circ D.$$

Take $T_2 \geq C_{4.3} K_1 \log n \log d$ and $K_2 \geq C_{4.6} \epsilon^{-2} \log n$. Sample Q'_{T_2} and let

$$\Psi_2 := \sqrt{\frac{K_1}{K_2}} \cdot \text{Sample}_{K_1, K_2} \circ Q'_{T_2} \circ D'.$$

Return

$$\Psi = \Psi_2 \circ \Psi_1.$$

#If ORA is replaced by S-ORA, then D, D' may be omitted and the first projection may be replaced with $\text{Proj}_{\text{Binom}(d, K_1/d)}$ and the second projection with Proj_{K_2} .

Lemma 4.2. *Let \mathbf{x} be an \mathbb{S}^{n-1} -valued random vector, and let $R = R_{i,j,\theta}$ be a random rotation corresponding to a single step of ORA. Then,*

$$\mathbb{E}_{R, \mathbf{x}}[S_k(R\mathbf{x})] \leq \left(1 - \frac{2}{d}\right) \mathbb{E}_{\mathbf{x}}[S_k(\mathbf{x})] + \frac{2^{1-k}}{d(d-1)} \sum_{a=0}^k \mathbb{E}_{\mathbf{x}}[S_a(\mathbf{x})] \mathbb{E}_{\mathbf{x}}[S_{k-a}(\mathbf{x})].$$

Proof. By direct computation using the definition of R , we have

$$\begin{aligned} \mathbb{E}_{R, \mathbf{x}}[S_k(R\mathbf{x})] &= \left(1 - \frac{2}{d}\right) \mathbb{E}_{\mathbf{x}}[S_k(\mathbf{x})] \\ &\quad + \frac{1}{(2k)! d(d-1)} \left(\sum_{i \neq j} \mathbb{E}_{\mathbf{x}} \left[\left(\frac{x_i + x_j}{\sqrt{2}} \right)^{2k} \right] + \sum_{i \neq j} \mathbb{E}_{\mathbf{x}} \left[\left(\frac{x_i - x_j}{\sqrt{2}} \right)^{2k} \right] \right) \\ &= \left(1 - \frac{2}{d}\right) \mathbb{E}_{\mathbf{x}}[S_k(\mathbf{x})] + \frac{1}{d(d-1)} \sum_{i \neq j} \sum_{a=0}^k \frac{2^{1-k}}{(2k)!} \binom{2k}{2a} \mathbb{E}_{\mathbf{x}}[x_i^{2a} x_j^{2k-2a}] \\ &\leq \left(1 - \frac{2}{d}\right) \mathbb{E}_{\mathbf{x}}[S_k(\mathbf{x})] + \frac{1}{d(d-1)} \sum_{i \neq j} \sum_{a=0}^k \frac{2^{1-k}}{(2k)!} \binom{2k}{2a} \mathbb{E}_{\mathbf{x}}[x_i^{2a}] \mathbb{E}_{\mathbf{x}}[x_j^{2k-2a}] \\ &\leq \left(1 - \frac{2}{d}\right) \mathbb{E}_{\mathbf{x}}[S_k(\mathbf{x})] + \frac{2^{1-k}}{d(d-1)} \sum_{a=0}^k \mathbb{E}_{\mathbf{x}}[S_a(\mathbf{x})] \mathbb{E}_{\mathbf{x}}[S_{k-a}(\mathbf{x})]. \end{aligned}$$

The first inequality follows from the fact that, conditioned on $\mathbf{x}_{-i,-j}$, (x_i^{2a}, x_j^{2k-2a}) is distributed as $(y^{2a}, (\sqrt{r^2 - y^2})^{2k-2a})$, where $r \geq 0$ is determined by $\mathbf{x}_{-i,-j}$, and y^2 is some distribution (determined by the original distribution on \mathbb{S}^{n-1} and $\mathbf{x}_{-i,-j}$) on the interval $[0, r^2]$. Since the first coordinate is a non-decreasing function of y^2 and the second coordinate is a non-increasing function of y^2 , it

follows from the FKG inequality that

$$\mathbb{E}[y^{2a}(r^2 - y^2)^{k-a}] \leq \mathbb{E}[y^{2a}]\mathbb{E}[(r^2 - y^2)^{k-a}]. \quad \square$$

From this lemma and a careful computation, one can deduce the following upper bound on the p -th moments of the coordinates of \mathbf{x} .

Proposition 4.3. *There exists an absolute constant $C_{4.3}$ for which the following holds. Let \mathbf{x} be an \mathbb{S}^{n-1} distributed random vector (in particular, \mathbf{x} can be deterministic). Fix a dimension $d \geq 25$, a positive integer $p \leq d$, and consider a time $t \geq C_{4.3}pd \log d$. Then,*

$$\mathbb{E}_{Q_t, \mathbf{x}}[S_p(Q_t \mathbf{x})] \leq \frac{2^{p-2}d^{1-p}}{p!}.$$

Remark. The proof below shows that taking $C_{4.3} = 2.25$ is sufficient.

Proof. We will prove this by strong induction on $p \geq 1$. Also, for lightness of notation, we will omit subscripts in the expectation.

For $p = 1$, note that $S_1(Q_t \mathbf{x}) = 1/2$ deterministically, so that the assertion holds. Hence, let $p \geq 2$, and suppose we know the statement for $1, \dots, p-1$. Let $e_{q,t} = \mathbb{E}[S_q(Q_t \mathbf{x})]$.

Let $t' = C_{4.3}(p-1)d \log d$, and note that $Q_t \mathbf{x} = Q_{t-t'}(Q_{t'} \mathbf{x}) \sim Q_{t-t'} \mathbf{y}$, where \mathbf{y} is an \mathbb{S}^{n-1} -valued random vector distributed as $Q_{t'} \mathbf{x}$. Hence, by the inductive hypothesis, we have that for all $t \geq t'$ and $0 \leq q \leq p-1$,

$$e_{q,t} = \mathbb{E}[S_q(Q_{t-t'} \mathbf{y})] \leq \frac{2^{q-2}d^{1-q}}{q!}.$$

Therefore, by Lemma 4.2 and the above, we have for $t \geq t'$ that

$$\begin{aligned} e_{p,t+1} &= \mathbb{E}[S_p(Q_{t+1-t'} \mathbf{y})] & (4.1) \\ &\leq \left(1 - \frac{2}{d}\right) \mathbb{E}[S_p(Q_{t-t'} \mathbf{y})] + \frac{2^{1-p}}{d(d-1)} \sum_{a=0}^p \mathbb{E}[S_a(Q_{t-t'} \mathbf{y})] \mathbb{E}[S_{p-a}(Q_{t-t'} \mathbf{y})] \\ &\leq \left(1 - \frac{2}{d} + \frac{2^{2-p}}{d-1}\right) \mathbb{E}[S_p(Q_{t-t'} \mathbf{y})] + \left(\frac{2^{1-p}}{d(d-1)} \sum_{a=1}^{p-1} \frac{2^{p-4}d^{2-p}}{a!(p-a)!}\right) \\ &\leq \left(1 - \frac{2}{d} + \frac{2^{2-p}}{d-1}\right) \mathbb{E}[S_p(Q_{t-t'} \mathbf{y})] + \left(\frac{2^{-3}}{d-1} \frac{d^{1-p}(2^p-2)}{p!}\right) \\ &= \left(1 - \frac{2}{d} + \frac{2^{2-p}}{d-1}\right) e_{p,t} + \left(\frac{2^{-3}}{d-1} \frac{d^{1-p}(2^p-2)}{p!}\right). & (4.2) \end{aligned}$$

To leverage the above relation, we also need to upper bound $e_{p,t'} = \mathbb{E}[S_p(\mathbf{y})]$. Indeed, by the inductive hypothesis, and the fact that each coordinate of \mathbf{y} is bounded in absolute value by 1, it follows that

$$e_{p,t'} = \mathbb{E}[S_p(\mathbf{y})] \leq \frac{1}{(2p)(2p-1)} \mathbb{E}[S_{p-1}(\mathbf{y})] \leq \frac{2^p \cdot d^{2-p}}{p!(2p-1)}. \quad (4.3)$$

To summarize, (4.2) and (4.3) demonstrate that

$$\begin{aligned} e_{p,t+1} &\leq \left(1 - \frac{2}{d} + \frac{2^{2-p}}{d-1}\right) e_{p,t} + \left(\frac{2^{-3}}{d-1} \frac{d^{1-p}(2^p-2)}{p!}\right). \\ e_{p,t'} &\leq \frac{2^p \cdot d^{2-p}}{p!(2p-1)}. \end{aligned}$$

Since $p \geq 2$ and $d \geq 25$, we have that

$$\left(1 - \frac{2}{d} + \frac{2^{2-p}}{d-1}\right) \leq \left(1 - \frac{2}{d} + \frac{1}{d-1}\right) \leq \left(1 - \frac{23}{24d}\right).$$

Therefore, by iterating the above relations, we have for $t > t'$ that

$$\begin{aligned} e_{p,t} &\leq \left(1 - \frac{23}{24d}\right)^{t-t'} \cdot \frac{2^p d^{2-p}}{p!(2p-1)} + \\ &\quad + \left(\frac{2^{-3}}{d-1} \frac{d^{1-p}(2^p-2)}{p!}\right) \sum_{j=0}^{t-t'-1} \left(1 - \frac{23}{24d}\right)^j \\ &\leq e^{-23(t-t')/24d} \cdot \frac{2^p d^{2-p}}{p!(2p-1)} + \frac{2^{p-3} d^{1-p}}{p!} \cdot \frac{25}{23}. \end{aligned}$$

In particular, for $t - t' \geq 48d \log d/23$, we see that

$$\begin{aligned} e_{p,t} &\leq \frac{2^{p-2} d^{1-p}}{p! \cdot 18} + \frac{2^{p-2} d^{1-p}}{p!} \cdot \frac{25}{46} \\ &\leq \frac{2^{p-2} d^{1-p}}{p!}, \end{aligned}$$

which completes the inductive step. \square

We will also need the following estimate regarding the maximum coordinate of $Q_t \mathbf{x}$; this estimate is better than simply applying Markov's inequality to [Proposition 4.3](#).

Proposition 4.4. *Fix a vector $\mathbf{x} \in \mathbb{S}^{d-1}$. Let $t \geq C_{4.4} d \log d \log \log d$. Then,*

$$\mathbb{P}\left[\|Q_t \mathbf{x}\|_\infty \geq 10\sqrt{\frac{\log d}{d}}\right] \leq d^{-2}.$$

Proof of Proposition 4.4. We may assume that $d \geq 10^2$ as otherwise, the desired conclusion holds trivially.

We will show the following: for any $p \in [d]$, there exists a collection of events A_1, \dots, A_p such that the following holds:

1. $A_1 \subseteq \dots \subseteq A_p$;
2. A_p depends only on the randomness used to generate the ORA for the first $C_{4.4}(d \log d \log(p-1) + d(p-1))$ steps;
3. $\mathbb{P}[A_j] \leq \frac{j}{d^5}$ for $1 \leq j \leq p$;
4. For any $t \geq C_{4.4}(d \log d \log p + dp)$,

$$\mathbb{E}[S_p(Q_t \mathbf{x}) | A_p^c] \leq 5^{p-1} \cdot \frac{d^{1-p}}{2^p \cdot p!}. \quad (4.4)$$

We prove this by strong induction on $p \geq 1$. For $p = 1$, we simply set $A_1 = \emptyset$ and note that $S_1(Q_t \mathbf{x}) = 1/2$ deterministically for all times t , so that the requirements for A_1 are trivially satisfied. Now suppose $p \geq 2$, and we know the statement for $0, \dots, p-1$.

Let B_p be the event that at time $t' = C_{4.4}(d \log d \log(p-1) + d(p-1))$, we have

$$S_{p-1}(Q_{t'}\mathbf{x}) \geq d^5 \cdot 5^{p-2} \frac{d^{2-p}}{2^{p-1}(p-1)!}.$$

Clearly, B_p only depends on the randomness used to generate the first $C_{4.4}d \log d \log(p-1)$ steps. Moreover, by Markov's inequality and the inductive hypothesis, we have that

$$\mathbb{P}[B_p | A_{p-1}^c] \leq \frac{1}{d^5}$$

and therefore, if we set $A_p = B_p \cup A_{p-1}$ then A_p satisfies the first three conclusions of the inductive hypothesis. To complete the inductive step, we only need to verify the last conclusion.

For this, we begin by noting that deterministically under A_p^c ,

$$\|Q_{t'}\mathbf{x}\|_\infty \leq L_p := 2\sqrt{2p - 2d^{\frac{7-p}{2(p-1)}}}.$$

The key feature of this bound that we need is that $L_p = \sqrt{p}d^{-1/2+\Theta(1/p)}$. Thus, by the induction hypothesis,

$$\begin{aligned} \mathbb{E}[S_p(Q_{t'}\mathbf{x}) | A_p^c] &\leq \frac{L_p^2}{(2p)(2p-1)} \mathbb{E}[S_{p-1}(Q_{t'}\mathbf{x}) | A_p^c] \\ &\leq \frac{L_p^2}{(2p)(2p-1)} \mathbb{E}[S_{p-1}(Q_{t'}\mathbf{x}) | A_{p-1}^c] \cdot \left(1 + \frac{1}{d^4}\right) \\ &\leq 5^{p+2} \frac{d^{1-p+6/(p-1)}}{2^p p!}. \end{aligned} \tag{4.5}$$

Let $e_{q,t} = \mathbb{E}[S_q(Q_t\mathbf{x}) | A_p^c]$. For $t \geq t'$ the distribution of $Q_t\mathbf{x}$ is the same as the distribution of $Q_{t-t'}\mathbf{y}$, where \mathbf{y} is an \mathbb{S}^{n-1} -valued random vector distributed as $Q_{t'}\mathbf{x}$. Also, by the inductive hypothesis, we have that for all $t \geq t'$ and $1 \leq q \leq p-1$,

$$e_{q,t} = \mathbb{E}[S_q(Q_{t-t'}\mathbf{y}) | A_p^c] \leq 5^{q-2} \frac{d^{1-q}}{2^q q!} \left(1 + \frac{1}{d^4}\right),$$

where, as before, the final factor comes from conditioning on A_p^c and not A_q^c . Therefore, by a trivial modification of [Lemma 4.2](#), we have for $t \geq t'$ that

$$\begin{aligned} e_{p,t+1} &= \mathbb{E}[S_p(Q_{t+1-t'}\mathbf{y}) | A_p^c] \leq \left(1 - \frac{2}{d}\right) \mathbb{E}[S_p(Q_{t-t'}\mathbf{y}) | A_p^c] \\ &\quad + \frac{2^{1-p}}{d(d-1)} \sum_{a=0}^p \mathbb{E}[S_a(Q_{t-t'}\mathbf{y}) | A_p^c] \mathbb{E}[S_{p-a}(Q_{t-t'}\mathbf{y}) | A_p^c] \\ &\leq \left(1 - \frac{2}{d} + \frac{2^{2-p}}{d-1}\right) \mathbb{E}[S_p(Q_{t-t'}\mathbf{y}) | A_p^c] \\ &\quad + 5^{p-2} \left(\frac{2^{1-p}}{d(d-1)} \sum_{a=1}^{p-1} \frac{d^{2-p}}{2^p a!(p-a)!}\right) \left(1 + \frac{1}{d^4}\right)^2 \\ &= \left(1 - \frac{2}{d} + \frac{2^{2-p}}{d-1}\right) \mathbb{E}[S_p(Q_{t-t'}\mathbf{y}) | A_p^c] \\ &\quad + 5^{p-2} \left(\frac{2^{1-p}}{d-1} \frac{d^{1-p}(2^p-2)}{2^p p!}\right) \left(1 + \frac{1}{d^4}\right)^2 \end{aligned}$$

$$= \left(1 - \frac{2}{d} + \frac{2^{2-p}}{d-1}\right) e_{p,t} + 5^{p-2} \left(\frac{2^{1-p}}{d-1} \frac{d^{1-p}(2^p-2)}{2^p p!}\right) \left(1 + \frac{1}{d^4}\right)^2. \quad (4.6)$$

To summarize, (4.5) and (4.6) demonstrate that

$$\begin{aligned} e_{p,t+1} &\leq \left(1 - \frac{2}{d} + \frac{2^{2-p}}{d-1}\right) e_{p,t} + 5^{p-2} \left(\frac{2^{1-p}}{d-1} \frac{d^{1-p}(2^p-2)}{2^p p!}\right) \left(1 + \frac{1}{d^4}\right)^2, \\ e_{p,t'} &\leq 5^{p+2} \frac{d^{1-p+6/(p-1)}}{2^p p!}. \end{aligned}$$

Now, a very similar computation to the one in the proof of Proposition 4.3 shows that for

$$t - t' \geq O(1) \left(\frac{d \log d}{p} + d\right),$$

$e_{p,t} \leq 5^{p-1} \cdot \frac{d^{1-p}}{2^p p!}$, which completes the inductive step.

The proof of the conclusion of Proposition 4.4 now follows easily. Indeed, take $p = 10 \log d$, and note that $\mathbb{P}[A_p] \leq \frac{1}{d^4}$ and that for $t' \geq C_{4.4} d \log d \log \log d$ Markov's inequality applied to (4.4) yields

$$\mathbb{P}\left[S_p(Q_{t'} \mathbf{x}) \leq d^4 \cdot 5^{p-1} \frac{d^{1-p}}{2^p p!} \middle| A_p^c\right] \leq \frac{1}{d^4}.$$

Trivial estimation based on $S_p(\mathbf{x}) \geq \|\mathbf{x}\|_\infty^p / (2p)!$ gives the desired result. \square

Finally, we prove an estimate which will be required in the second phase of Algorithm 2. For this, we will make use of the following result of Latała [25].

Lemma 4.5 ([25, Corollary 2]). *For a random variable X , let $\|X\|_s = (\mathbb{E}|X|^s)^{1/s}$. There exists an absolute constant $C_{4.5}$ for which the following holds. Let X_1, \dots, X_n be independent copies of a symmetric random variable X . Then,*

$$\|X_1 + \dots + X_n\|_p \leq C_{4.5} \sup \left\{ \frac{p}{s} \left(\frac{n}{p}\right)^{1/s} \|X\|_s : \max(2, p/n) \leq s \leq p \right\}.$$

Lemma 4.6. *Let Q_t denote ORA of length t , and let $X_t = Q_t X_0$ with $t \geq C_{4.3} d \log d \log n$. Choose a uniformly random set S of indices of size $|S| = k$. If $k = C_{4.6} \epsilon^{-2} \log n$, then*

$$\mathbb{P}\left[\sum_{i \in S} X_t[i]^2 \notin \frac{k}{d} [1 - \epsilon, 1 + \epsilon]\right] \leq n^{-3}.$$

Proof. Choose k independent random indices i_1, \dots, i_k , potentially repeated. We first show that for any $p \geq 1$,

$$\mathbb{E} \left| \sum_{i \in S} X_t[i]^2 - \frac{k}{d} \right|^p \leq \mathbb{E} \left| \sum_{j=1}^k X_t[i_j]^2 - \frac{k}{d} \right|^p. \quad (4.7)$$

To see this, consider the joint distribution on $[d]^k \times \binom{[d]}{k}$ given by (i_1, \dots, i_k, T) , where i_1, \dots, i_k are independent random indices, potentially repeated, and T is a set of size k , chosen uniformly at random from among all subsets of $[d]$ of size k containing $\{i_1, \dots, i_k\}$. Note in particular that by

symmetry, the marginal distribution of T is uniform on $\binom{[d]}{k}$. Therefore, (4.7) will follow from the law of total probability if we can show that

$$\left| \sum_{i \in S} X_t[i]^2 - \frac{k}{d} \right|^p \leq \mathbb{E} \left[\left| \sum_{j=1}^k X_t[i_j]^2 - \frac{k}{d} \right|^p \middle| T = S \right]$$

for all $|S| = k$. But now, notice that the distribution on (i_1, \dots, i_k) conditioned on $T = S$ is *some* distribution on S^k which is symmetric under permutations of S . Thus, Jensen's inequality immediately implies (4.7).

For the remainder of the proof, we will focus on the model with k independent random indices. Let $Y_t = Q_t' Y_0$, where $Y_0 = X_0$ and Q_t' is an independent copy of Q_t . We have

$$\begin{aligned} \left(\mathbb{E} \left| \sum_{j=1}^k X_t[i_j]^2 - \frac{k}{d} \right|^p \right)^{1/p} &\leq \left(\mathbb{E} \left| \sum_{j=1}^k X_t[i_j]^2 - Y_t[i_j]^2 \right|^p \right)^{1/p} \\ &\leq C_{4.5} \sup_{2 \leq s \leq p} \frac{p}{s} \left(\frac{k}{p} \right)^{1/s} (\mathbb{E} |X_t[i_1]^2 - Y_t[i_1]^2|^s)^{1/s} \\ &\leq 2C_{4.5} \sup_{2 \leq s \leq p} \frac{p}{s} \left(\frac{k}{p} \right)^{1/s} (\mathbb{E}_{\{Q_t, i_1\}} |X_t[i_1]|^{2s})^{1/s} \\ &= 2C_{4.5} \sup_{2 \leq s \leq p} \frac{p}{s} \left(\frac{k}{p} \right)^{1/s} \left(\mathbb{E}_{Q_t} \frac{1}{d} \sum_{i=1}^d |X_t[i]|^{2s} \right)^{1/s}, \end{aligned}$$

where the first line uses Jensen's inequality, the second line uses Lemma 4.5, and the third line uses the triangle inequality.

By Proposition 4.3, if $1 \leq s \leq p$ is an integer, then

$$\mathbb{E}_{Q_t} \frac{1}{d} \sum_{i=1}^d |X_t[i]|^{2s} \leq \frac{2^{s-2} (2s)! d^{1-s}}{s!}$$

as long as $t \geq C_{4.3} p d \log d$. This (combined with Hölder's inequality to interpolate non-integer moments) shows that

$$\left(\mathbb{E} \left| \sum_{j=1}^k X_t[i_j]^2 - \frac{k}{d} \right|^p \right)^{1/p} \leq \left(\mathbb{E} \left| \sum_{j=1}^k X_t[i_j]^2 - Y_t[i_j]^2 \right|^p \right)^{1/p} \leq 2C_{4.5} \sup_{2 \leq s \leq p} \frac{p}{s} \left(\frac{k}{p} \right)^{1/s} \frac{10s}{d}.$$

Now (4.7) gives

$$\left(\mathbb{E} \left| \sum_{i \in S} X_t[i]^2 - \frac{k}{d} \right|^p \right)^{1/p} \leq 2C_{4.5} \sup_{2 \leq s \leq p} \frac{p}{s} \left(\frac{k}{p} \right)^{1/s} \frac{10s}{d}.$$

Now, for $k = C\epsilon^{-2} \log n$ and $p = \log n$, we see that the supremum is attained at $s = 2$, so that by Markov's inequality,

$$\mathbb{P} \left[\sum_{i \in S} X_t[i]^2 \notin \frac{k}{d} [1 - \epsilon, 1 + \epsilon] \right] \leq \left(\left(\frac{d}{k\epsilon} \right) \cdot 20C_{4.5} p \left(\frac{k}{p} \right)^{1/2} \frac{1}{d} \right)^p.$$

Choosing $C > 10^6 C_{4.5}^2$, we find that this is less than $1/n^3$, as desired. \square

We are now ready to prove [Theorem 1.6](#).

Proof of [Theorem 1.6](#). Let $q = K_1/d$. Applying [Theorem 3.4](#) and using [Proposition 4.4](#) at time $t = T_1 = C_{4.4}d \log d \log \log d$, we see that

$$\widetilde{\Psi}_1 = \frac{1}{\sqrt{q}} \text{Proj}_{d,q} \circ Q_{T_1},$$

with probability $1 - O(1/d)$, satisfies $\mathbb{P}[\delta_s(\widetilde{\Psi}_1) \geq \epsilon/4] \leq 1/d$ as long as

$$K_1 \gtrsim s(\log d)\epsilon^{-2}(\log s)^2(\log K_1)(\log d).$$

Note that in the case $K_1 = d$, the operator $\widetilde{\Psi}_1$ is actually orthogonal.

Now by [Theorem 3.5](#), we have that if $s \geq 40 \log(4n/\eta)$, then Ψ_1 acts as a $(1 \pm \epsilon)$ -isometry on our set of points X with probability at least $1 - \eta$. Choosing $\eta = 1/4$ and $s = 40 \log(16n)$, we see that this property holds with probability at least $3/4 - O(1/d)$ as long as

$$K_1 \gtrsim \epsilon^{-2}(\log n)(\log d)^2(\log \log n)^2(\log K_1).$$

Since $K_1 \leq d$,

$$K_1 \gtrsim \epsilon^{-2}(\log n)(\log d)^3(\log \log n)^2$$

certainly suffices. This indeed holds based on the choice of K_1 in [Algorithm 2](#), as long as C_2 is chosen large enough. Note that if we use S-ORA instead of ORA, then by [Lemma 2.3](#), this holds also for $\widetilde{\Psi}_1$, so that indeed, the random diagonal Rademacher matrix D may be excluded. Furthermore, due to the permutation symmetry in S-ORA established by [Lemma 2.4](#), we can replace $\text{Proj}_{d,q}$ in the definition of $\widetilde{\Psi}_1$ by $\text{Proj}_{\text{Binom}(d,q)}$, similar to the argument in the proof of [Theorem 1.4](#) (the symmetrization to Ψ_2 is similar and we will not further elaborate on this point).

By Markov's inequality, the probability that the actual number of dimensions in the image of $\text{Proj}_{d,q}$ is more than a constant times K_1 is sufficiently small, so with probability at least $5/7$ we have that $\widetilde{\Psi}_1$ is a $(1 \pm \epsilon)$ -isometry on our points and projects down to at most $O(\epsilon^{-2}(\log n)(\log d)^3(\log \log n)^2)$ dimensions.

To finish, we claim that Ψ_2 is a $(1 \pm \epsilon)$ -isometry on the image of our point set, $\Psi_1 X$ – as long as C_2 is large enough, this follows immediately by using [Lemma 4.6](#) and taking the union bound over all n vectors in the image. Since Ψ_1, Ψ_2 are both $(1 \pm \epsilon)$ -isometries on the relevant sets of points, we are immediately done (after rescaling ϵ): the desired isometry property holds with probability at least, say, $2/3$.

Finally, the analysis of the running time and memory of [Algorithm 2](#) is essentially identical to that of [Algorithm 1](#). \square

5 Open Problems

The most immediate problem left open by our work is to remove the additional $\log \log d$ term from [Theorem 1.6](#), and bring the ORA-based [Algorithm 2](#) on par with Kac walk and Hadamard matrix based transforms. Another intriguing question is whether algorithms based on the Kac walk can be used to successfully design optimal JL transforms beyond (1.4)/ optimal RIP transforms beyond (1.6), running in time $O(d \log d)$; indeed, the appearance of the error term $O(d \log n)$ in our bounds (as opposed to $O(\epsilon^{-2}d \log n)$) provides evidence that Kac walk based transforms outperform Hadamard matrix based transforms in large-data/high-accuracy regimes. Finally, it would be very interesting to compare how implementations of Kac walk or ORA-based transforms (optimized for issues/features such as cache locality, parallelization, and memory efficiency) compare to transforms based on Hadamard matrices; see [14] for some experimental results in this direction.

6 Acknowledgements

We thank Haim Avron and Sourav Chatterjee for helpful comments on an early version of this paper.

References

- [1] Dimitris Achlioptas, *Database-friendly random projections: Johnson-Lindenstrauss with binary coins*, Journal of computer and System Sciences **66** (2003), 671–687.
- [2] Nir Ailon and Bernard Chazelle, *The fast Johnson–Lindenstrauss transform and approximate nearest neighbors*, SIAM Journal on computing **39** (2009), 302–322.
- [3] Nir Ailon and Edo Liberty, *Fast dimension reduction using Rademacher series on dual BCH codes*, Discrete & Computational Geometry **42** (2009), 615.
- [4] Nir Ailon and Holger Rauhut, *Fast and RIP-optimal transforms*, Discrete Comput. Geom. **52** (2014), 780–798.
- [5] Noga Alon and Bo’az Klartag, *Optimal compression of approximate inner products and dimension reduction*, 2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS), IEEE, 2017, pp. 639–650.
- [6] Haim Avron, Petar Maymounkov, and Sivan Toledo, *Blendenpik: Supercharging LAPACK’s least-squares solver*, SIAM Journal on Scientific Computing **32** (2010), 1217–1236.
- [7] Keith Ball, *An elementary introduction to modern convex geometry*, Flavors of geometry, Math. Sci. Res. Inst. Publ., vol. 31, Cambridge Univ. Press, Cambridge, 1997, pp. 1–58.
- [8] Stefan Bamberger and Felix Krahmer, *Optimal fast Johnson-Lindenstrauss embeddings for large data sets*, arXiv:1712.01774.
- [9] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin, *A simple proof of the restricted isometry property for random matrices*, Constructive Approximation **28** (2008), 253–263.
- [10] Béla Bollobás, *Random graphs*, second ed., Cambridge Studies in Advanced Mathematics, vol. 73, Cambridge University Press, Cambridge, 2001.
- [11] Emmanuel J Candès, Justin Romberg, and Terence Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, IEEE Transactions on information theory **52** (2006), 489–509.
- [12] Emmanuel J. Candes and Terence Tao, *Near-optimal signal recovery from random projections: universal encoding strategies?*, IEEE Trans. Inform. Theory **52** (2006), 5406–5425.
- [13] Sourav Chatterjee, Persi Diaconis, Allan Sly, and Lingfu Zhang, *A phase transition for repeated averages*, 2019.
- [14] Krzysztof Choromanski, Mark Rowland, Wenyu Chen, and Adrian Weller, *Unifying orthogonal Monte Carlo methods*, International Conference on Machine Learning, 2019, pp. 1203–1212.
- [15] Sanjoy Dasgupta and Anupam Gupta, *An elementary proof of a theorem of Johnson and Lindenstrauss*, Random Structures & Algorithms **22** (2003), 60–65.

- [16] Persi Diaconis and Mehrdad Shahshahani, *Generating a random permutation with random transpositions*, Z. Wahrsch. Verw. Gebiete **57** (1981), 159–179.
- [17] Sjoerd Dirksen, *Tail bounds via generic chaining*, Electron. J. Probab. **20** (2015), no. 53, 29.
- [18] David L Donoho, *Compressed sensing*, IEEE Transactions on information theory **52** (2006), 1289–1306.
- [19] Ishay Haviv and Oded Regev, *The restricted isometry property of subsampled Fourier matrices*, Geometric Aspects of Functional Analysis, Springer, 2017, pp. 163–179.
- [20] William B Johnson and Joram Lindenstrauss, *Extensions of Lipschitz mappings into a Hilbert space*, Contemporary mathematics **26** (1984), 1.
- [21] Mark Kac, *Foundations of kinetic theory*, Proceedings of The third Berkeley symposium on mathematical statistics and probability, vol. 3, University of California Press Berkeley and Los Angeles, California, 1956, pp. 171–197.
- [22] Daniel M Kane and Jelani Nelson, *Sparser Johnson-Lindenstrauss transforms*, Journal of the ACM (JACM) **61** (2014), 1–23.
- [23] Felix Krahmer and Rachel Ward, *New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property*, SIAM J. Math. Anal. **43** (2011), 1269–1281.
- [24] Kasper Green Larsen and Jelani Nelson, *Optimality of the Johnson–Lindenstrauss lemma*, 2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS), IEEE, 2017, pp. 633–638.
- [25] Rafał Łatała, *Estimation of moments of sums of independent real random variables*, Ann. Probab. **25** (1997), 1502–1513.
- [26] Roberto Imbuzeiro Oliveira et al., *On the convergence to equilibrium of kac’s random walk on matrices*, The Annals of Applied Probability **19** (2009), 1200–1231.
- [27] Natesh S. Pillai and Aaron Smith, *Kac’s walk on n -sphere mixes in $n \log n$ steps*, Ann. Appl. Probab. **27** (2017), 631–650.
- [28] Natesh S. Pillai and Aaron Smith, *On the mixing time of Kac’s walk and other high-dimensional Gibbs samplers with constraints*, Ann. Probab. **46** (2018), 2345–2399.
- [29] Holger Rauhut, *Compressive sensing and structured random matrices*, Theoretical foundations and numerical methods for sparse recovery, Radon Ser. Comput. Appl. Math., vol. 9, Walter de Gruyter, Berlin, 2010, pp. 1–92.
- [30] Mark Rudelson and Roman Vershynin, *On sparse reconstruction from Fourier and Gaussian measurements*, Comm. Pure Appl. Math. **61** (2008), 1025–1045.

A Proof of Lemma 3.2

Proof. Let $A_t[i] = X_t[i]^2$ and $B_t[i] = Y_t[i]^2$ for all $t \geq 0$ and $i \in [d]$ and recall that X_t and Y_t are coupled as in Definition 3.1. We calculate,

$$\mathbb{E} \left[\sum_{k=1}^d (A_1[k] - B_1[k])^2 \right] = \frac{2}{d(d-1)} \sum_{1 \leq i < j \leq d} \mathbb{E} \left[\sum_{k=1}^d (A_1[k] - B_1[k])^2 | (i_0, j_0) = (i, j) \right]$$

$$\begin{aligned}
&= \frac{2}{d(d-1)} \frac{(d-1)(d-2)}{2} \sum_{k=1}^d (A_0[k] - B_0[k])^2 \\
&\quad + \frac{2}{d(d-1)} \sum_{i<j} \mathbb{E} \left[\left((A_0[i] + A_0[j]) \cos^2 \varphi - (B_0[i] + B_0[j]) \cos^2 \varphi \right)^2 \right] \\
&\quad + \frac{2}{d(d-1)} \sum_{i<j} \mathbb{E} \left[\left((A_0[i] + A_0[j]) \sin^2 \varphi - (B_0[i] + B_0[j]) \sin^2 \varphi \right)^2 \right] \\
&= \frac{d-2}{d} \sum_{k=1}^d (A_0[k] - B_0[k])^2 \\
&\quad + \frac{4}{d(d-1)} \mathbb{E}[\cos^4 \varphi] \sum_{i<j} \left((A_0[i] + A_0[j]) - (B_0[i] + B_0[j]) \right)^2 \\
&= \left(1 - \frac{2}{d}\right) \sum_{k=1}^d (A_0[k] - B_0[k])^2 + \frac{3}{2d(d-1)} \sum_{i<j} \left((A_0[i] + A_0[j]) - (B_0[i] + B_0[j]) \right)^2 \\
&= \left(1 - \frac{2}{d}\right) \sum_{k=1}^d (A_0[k] - B_0[k])^2 + \frac{3}{2d(d-1)} \sum_{i<j} \left((A_0[i] - B_0[i])^2 + (A_0[j] - B_0[j])^2 \right) \\
&\quad + \frac{3}{d(d-1)} \sum_{i<j} (A_0[i] - B_0[i])(A_0[j] - B_0[j]) \\
&= \left(1 - \frac{2}{d}\right) \sum_{k=1}^d (A_0[k] - B_0[k])^2 + \frac{3}{2d} \sum_{k=1}^d (A_0[k] - B_0[k])^2 \\
&\quad + \frac{3}{d(d-1)} \sum_{i<j} (A_0[i] - B_0[i])(A_0[j] - B_0[j]) \\
&= \left(1 - \frac{1}{2d}\right) \sum_{k=1}^d (A_0[k] - B_0[k])^2 + \frac{3}{d(d-1)} \sum_{i<j} (A_0[i] - B_0[i])(A_0[j] - B_0[j]) \\
&= \left(1 - \frac{1}{2d}\right) \sum_{k=1}^d (A_0[k] - B_0[k])^2 \\
&\quad + \frac{3}{2d(d-1)} \left(\left(\sum_{k=1}^d (A_0[k] - B_0[k]) \right)^2 - \sum_{k=1}^d (A_0[k] - B_0[k])^2 \right) \\
&= \left(1 - \frac{1}{2d} - \frac{3}{2d(d-1)}\right) \sum_{k=1}^d (A_0[k] - B_0[k])^2,
\end{aligned}$$

where the last equality uses $\sum_{k=1}^d A_0[k] = 1 = \sum_{k=1}^d B_0[k]$. Thus, we have

$$\mathbb{E} \left[\sum_{k=1}^d (A_1[k] - B_1[k])^2 \right] \leq \left(1 - \frac{1}{2d}\right).$$

For $t \geq 0$, let \mathcal{F}_t denote the σ -algebra generated by the random variables X_0, \dots, X_t and Y_0, \dots, Y_t . Repeatedly applying the previous inequality, we have for all $t \geq 0$ that

$$\begin{aligned}
\mathbb{E} \left[\sum_{k=1}^d (A_t[k] - B_t[k])^2 \right] &= \mathbb{E} \left[\mathbb{E} \left[\sum_{k=1}^d (A_t[k] - B_t[k])^2 \mid \mathcal{F}_{t-1} \right] \right] \\
&\leq \left(1 - \frac{1}{2d}\right) \mathbb{E} \left[\sum_{k=1}^d (A_{t-1}[k] - B_{t-1}[k])^2 \right]
\end{aligned}$$

$$\begin{aligned}
&\leq \left(1 - \frac{1}{2d}\right)^t \sum_{k=1}^d \mathbb{E} [(A_0[k] - B_0[k])^2] \\
&\leq 2 \left(1 - \frac{1}{2d}\right)^t
\end{aligned}$$

as desired. \square

B Proof of Lemma 2.4

Throughout this section, we will freely use various tools from nonabelian Fourier analysis; we refer the reader to [16] for an introduction to such techniques. We will let ρ denote an irreducible representation of \mathfrak{S}_d , i.e., $\rho \in \widehat{\mathfrak{S}}_d$ and d_ρ denote its dimension. Since \mathfrak{S}_d is finite, all its finite dimensional representations are unitarizable, and we will work with a choice of inner product such that irreducible representations are also unitary. In particular, various appearances of \dagger should be understood as the operator-theoretic adjoint with respect to the appropriate inner product. A sum over nontrivial irreducible representations will be denoted \sum'_ρ . The key estimate we need is the following purely probabilistic claim regarding permutations.

Lemma B.1. *Let $\sigma = (\sigma_1, \dots, \sigma_T)$ be uniformly randomly chosen transpositions in \mathfrak{S}_d and suppose $d \geq 10$. Let $\xi_i \sim \text{Ber}(1/2)$ for $1 \leq i \leq T$. Define*

$$P_\sigma = \sigma_1^{\xi_1} \cdots \sigma_T^{\xi_T}.$$

Then

$$\mathbb{E}_{\sigma, \xi} [\text{TV}(P_\sigma, \text{Unif}_{\mathfrak{S}_d})] \leq C_{B.1} \left(d^{1/2} e^{-T/(6d)} + (d!)^{1/2} \left(\frac{\sqrt{5}-1}{2} \right)^{T/2} \right)$$

for an absolute constant $C_{B.1} > 0$.

Remark. The given proof can be modified (with more careful character estimates similar to [16]) to show the quantity studied tends to 0 once T passes $2d \log d$ (with a $\Theta(d)$ rate). It is an interesting question as to whether this is the sharp cutoff.

Proof. Let $U : \mathfrak{S}_d \rightarrow \mathbb{C}$ be $1/d!$ everywhere. For any permutation τ , let $f_\tau : \mathfrak{S}_d \rightarrow \mathbb{C}$ be $1/2$ at the identity and τ , and 0 elsewhere. We note that $\widehat{U}(\rho) = 0$ for nontrivial representations ρ . We also note that

$$\mathbb{P}[P_\sigma = \tau] = f_{\sigma_1} * \cdots * f_{\sigma_T}(\tau)$$

by the definition of convolution. The Fourier coefficient of this function at ρ is

$$A_\sigma(\rho) := \widehat{f_{\sigma_1}}(\rho) \cdots \widehat{f_{\sigma_T}}(\rho).$$

By the proof of the upper bound lemma of Diaconis and Shahshahani [16], we have

$$\begin{aligned}
\mathbb{E}_\sigma [\text{TV}(P_\sigma, \text{Unif}_{\mathfrak{S}_d})] &= \mathbb{E}_\sigma \left[\sum_{\tau \in \mathfrak{S}_d} \left| \mathbb{P}[P_\sigma = \tau] - \frac{1}{d!} \right| \right] \\
&\leq \left[\mathbb{E}_\sigma \left[d! \sum_{\tau \in \mathfrak{S}_d} \left(\mathbb{P}[P_\sigma = \tau] - \frac{1}{d!} \right)^2 \right] \right]^{1/2}
\end{aligned}$$

$$= \left[\mathbb{E}_\sigma \left[\sum'_\rho d_\rho \operatorname{Tr}(A_\sigma(\rho)A_\sigma(\rho)^\dagger) \right] \right]^{1/2},$$

where the first line is by definition, the second line uses Cauchy–Schwarz, and the third line uses Plancherel’s formula. In the third line, we also used that U has zero Fourier coefficient at nontrivial representations, and that the term at the trivial representation cancels out.

Next, we claim that

$$\mathbb{E}_{\sigma_i} \left[\widehat{f_{\sigma_i}}(\rho) \widehat{f_{\sigma_i}}(\rho)^\dagger \right] = c_\rho I_{d_\rho}$$

for a constant $c_\rho \in \mathbb{R}$. In fact, we can compute this constant explicitly. Let χ_ρ be the trace of ρ evaluated at any transposition, and let $r_\rho = \chi_\rho/d_\rho$. It is worth noting that $|r_\rho| \leq 1$ since unitary matrices have trace at most d_ρ . We find, since $\rho(\sigma_i)$ is unitary, that

$$\mathbb{E}_{\sigma_i} \left[\widehat{f_{\sigma_i}}(\rho) \widehat{f_{\sigma_i}}(\rho)^\dagger \right] = \mathbb{E}_{\sigma_i} \left[\left(\frac{I_{d_\rho} + \rho(\sigma_i)}{2} \right) \left(\frac{I_{d_\rho} + \rho(\sigma_i)}{2} \right)^\dagger \right] = \frac{1}{2} I_{d_\rho} + \frac{1}{2} \mathbb{E}_{\sigma_i} [\rho(\sigma_i)] = \left(\frac{1 + r_\rho}{2} \right) I_{d_\rho}.$$

In the last step we noted that $\mathbb{E}_{\sigma_i}[\rho(\sigma_i)]$ is a multiple of the identity by Schur’s lemma (or, it is the Fourier transform of a function constant on conjugacy classes) and has trace χ_ρ by definition (note that χ_ρ is real since $\rho(\sigma_i)$ is an involution). Thus

$$c_\rho = \frac{1 + r_\rho}{2}.$$

Now, note that

$$\begin{aligned} \mathbb{E}_\sigma \left[\operatorname{Tr}(A_\sigma(\rho)A_\sigma(\rho)^\dagger) \right] &= \mathbb{E}_\sigma \left[\operatorname{Tr} \left(\widehat{f_{\sigma_1}}(\rho) \cdots \widehat{f_{\sigma_T}}(\rho) \widehat{f_{\sigma_T}}(\rho)^\dagger \cdots \widehat{f_{\sigma_1}}(\rho)^\dagger \right) \right] \\ &= c_\rho \mathbb{E}_{\sigma_1, \dots, \sigma_{T-1}} \left[\operatorname{Tr} \left(\widehat{f_{\sigma_1}}(\rho) \cdots \widehat{f_{\sigma_{T-1}}}(\rho) \widehat{f_{\sigma_{T-1}}}(\rho)^\dagger \cdots \widehat{f_{\sigma_1}}(\rho)^\dagger \right) \right] \\ &= \dots \\ &= d_\rho c_\rho^T = d_\rho \left(\frac{1 + r_\rho}{2} \right)^T. \end{aligned}$$

Therefore

$$\mathbb{E}_\sigma [\operatorname{TV}(P_\sigma, \operatorname{Unif}_{\mathfrak{S}_d})] \leq \left[\sum'_\rho d_\rho^2 \left(\frac{1 + r_\rho}{2} \right)^T \right]^{1/2},$$

and it remains to bound the right side.

The key technical result in [16, p. 27] is that

$$\left[\sum'_\rho d_\rho^2 \left(\frac{1}{d} + \frac{d-1}{d} r_\rho \right)^{2k} \right]^{1/2} \leq C d e^{-2k/d}$$

where C is an absolute constant independent of d . Now if $r_\rho \in [\sqrt{5} - 2, 1]$ we have

$$0 < \left(\frac{1 + r_\rho}{2} \right)^3 \leq r_\rho \leq \frac{1}{d} + \frac{d-1}{d} r_\rho,$$

while if $r_\rho \in [-1, \sqrt{5} - 2]$ we have

$$\left| \frac{1 + r_\rho}{2} \right| \leq \frac{\sqrt{5} - 1}{2}.$$

Thus, using this, we see that if $6|T$ we have

$$\begin{aligned} \sum_{\rho}' d_{\rho}^2 \left(\frac{1+r_{\rho}}{2} \right)^T &\leq \sum_{\rho}' d_{\rho}^2 \left(\frac{1}{d} + \frac{d-1}{d} r_{\rho} \right)^{T/3} + \sum_{\rho}' d_{\rho}^2 \left(\frac{\sqrt{5}-1}{2} \right)^T \\ &\leq Cde^{-T/(3d)} + d! \left(\frac{\sqrt{5}-1}{2} \right)^T. \end{aligned}$$

Since the TV is decreasing as T increases, the result follows immediately by rounding T to the nearest multiple of 6. \square

Now we are ready to prove [Lemma 2.4](#).

Proof of Lemma 2.4. By two applications of [Lemma 2.3](#), we see that

$$\text{TV}(Q_T, D_{\xi} Q_T D_{\xi'}) \leq \frac{2d \exp\left(-\frac{T}{d-1}\right)}{1 - d \exp\left(-\frac{T}{d-1}\right)} \quad (\text{B.1})$$

if ξ, ξ' are independent random vectors which are uniform over $\{\pm 1\}^d$, conditioned on having product 1.

Now, let $Q_T = R_{i_T, j_T, \theta_T} \cdots R_{i_1, j_1, \theta_1}$ as usual. For every pair of distinct indices $i, j \in [d]$, let $D_{i,j}$ be the random rotation in the (i, j) plane by a uniform multiple of $\pi/2$. For every time $t \in [T]$, let D_t be a random matrix distributed as D_{i_t, j_t} , sampled independently from everything except (i_t, j_t) . First, note that R_{i_t, j_t, θ_t} and $R_{i_t, j_t, \theta_t} D_t$ have the same distribution since our distribution q on angles is invariant under $\theta \leftrightarrow \theta + k\pi/2$ for all $k \in \mathbb{Z}$. Second, note that the distributions

$$D_{i', j'} R_{i, j, \theta} \text{ and } R_{i, j, \theta} D_{i', j'}$$

are the same. The reason is more subtle than in the proof of [Lemma 2.3](#). The point is that $D_{i', j'}$ merely permutes and signs the basis vectors e_1, \dots, e_d (via at worst a transposition). Thus conjugation of $R_{i, j, \theta}$ by $D_{i', j'}$ gives another rotation in a coordinate plane $(\sigma(i), \sigma(j))$ (where σ is either the identity or the swap $(i' j')$), with its angle potentially changed via negation, addition by π , or both. Either way, we see $D_{i', j'} R_{i, j, \theta} D_{i', j'}^{-1}$ (conditional on the value $D_{i', j'}$) has the same distribution as $R_{i, j, \theta}$, hence the claim.

Now we extract the matrices D_t similar to in the proof of [Lemma 2.3](#). However, we must be slightly careful: note that D_t is dependent on (i_t, j_t) , and the swapping operation above can potentially change a pair (i_t, j_t) as we move past (which was not true before). Therefore, we will perform swaps in a way such that once D_t has been extracted to the end, the rotation R_{i_t, j_t, θ_t} is not touched again. In fact, we were careful to do this already in the proof of [Lemma 2.3](#), although this care was not needed there.

Specifically, we apply the first operation to R_{i_T, j_T, θ_T} , and then apply the second operation repeatedly to switch the diagonal matrix D_T to the end. Then we do the same for $R_{i_{T-1}, j_{T-1}, \theta_{T-1}}$, and so on. We thus see that Q_T has the same distribution as

$$Q_T D_1 \cdots D_T.$$

Let $\sigma_t = (i_t j_t)$ for $1 \leq t \leq T$. Note that $D_1 \cdots D_T$ is independent of Q_T conditional on $\sigma = (\sigma_1, \dots, \sigma_T)$, and is a signed permutation matrix with determinant 1. Therefore it can be written uniquely as $D_1 \cdots D_T = PD$, where P is an unsigned permutation matrix and D is a diagonal sign matrix, with $\det(PD) = 1$. Note that (P, D) is independent of Q_T conditional on σ .

Furthermore, we see that we can change (P, D) into a joint distribution on signed permutation matrices (with determinant 1) and diagonal matrices (with determinant 1) which has a uniform marginal on P while sacrificing at most a TV of

$$\text{TV}(P_\sigma, \text{Unif}_{\mathfrak{S}_d}),$$

where P_σ is defined as in [Lemma B.1](#). This is since (conditional on σ) $D_1 \cdots D_T$ induces a permutation on the coordinates e_1, \dots, e_d with the same distribution as P_σ .

Let Σ be a uniform signed permutation matrix with determinant 1. We deduce that there is a distribution of diagonal matrices D (with determinant 1), potentially dependent on $(i_1, j_1), \dots, (i_T, j_T)$ and Σ , such that

$$\text{TV}(Q_T D_1 \cdots D_T, Q_T \Sigma D) \leq \mathbb{E}_\sigma \text{TV}(P_\sigma, \text{Unif}_{\mathfrak{S}_d}).$$

Therefore, for $D_\xi, D_{\xi'}$ independent from everything as defined at the beginning, we have

$$\text{TV}(D_\xi Q_T D_{\xi'}, D_\xi Q_T \Sigma D D_{\xi'}) \leq \mathbb{E}_\sigma \text{TV}(P_\sigma, \text{Unif}_{\mathfrak{S}_d}).$$

Regardless of the value of D , we see that the independent sign matrix $D_{\xi'}$ rerandomizes it so that $D_\xi Q_T \Sigma D D_{\xi'}$ and $D_\xi Q_T \Sigma$ have the same distribution. Using this, along with [\(B.1\)](#), we deduce that

$$\text{TV}(Q_T, D_\xi Q_T \Sigma) \leq \frac{2d \exp\left(-\frac{T}{d-1}\right)}{1 - d \exp\left(-\frac{T}{d-1}\right)} + \mathbb{E}_\sigma \text{TV}(P_\sigma, \text{Unif}_{\mathfrak{S}_d}),$$

and now [Lemma B.1](#) finishes. Technically, we also note that Q_T is invariant under taking transposes, so that we can also deduce a bound on $\text{TV}(Q_T, \Sigma Q_T D_\xi)$. \square